CLAIMS

WE CLAIM:

1. A machine-readable medium having stored thereon executable instructions that when executed by a processor, cause the processor to:

generate frequency vectors for each non-context token in a corpus based upon counted occurrences of a predetermined relationship of the non-context tokens to context tokens; and

cluster the non-context tokens into a cluster tree based upon the frequency vectors according to a lexical correlation among the non-context tokens.

2. A method of grammar learning from a corpus, comprising:

generating frequency vectors for each non-context token in a corpus based upon counted occurrences of a predetermined relationship of the non-context tokens to context tokens; and

clustering the non-context tokens based upon the frequency vectors according to a lexical correlation among the non-context tokens.

3. The method of claim 2, wherein the step of clustering further comprises clustering the non-context tokens into a cluster tree.

4. The method of claim 3, wherein the cluster tree represents a grammatical relationship among the non-context tokens.

5. The method of claim 3, further comprising cutting the cluster tree along a cutting line to separate large clusters from small clusters.

6. The method of claim 2, wherein small clusters are ranked according to a compactness value.

7. The method of claim 2, wherein the predetermined relationship is a measure of adjacency.

8. The method of claim 2, wherein the clustering is performed based on Euclidean distances between the frequency vectors.

9. The method of claim 2, wherein the clustering is performed based on Manhattan distances between the frequency vectors.

10. The method of claim 2, wherein the clustering is performed based on maximum distance metrics between the frequency vectors.

11. The method of claim 2, further comprising normalizing the frequency vectors based upon a number of occurrences of the non-context token in the corpus.

12. The method of claim 2, wherein the frequency vectors are multi-dimensional vectors, the number of dimensions being determined by the number of context tokens and a number of predetermined relationships of non-context tokens to the context token being counted.

13. A file storing a grammar model of a corpus of speech, created according to a method comprising:

generating frequency vectors for each non-context token in a corpus based upon counted occurrences of a predetermined relationship of the non-context tokens to context tokens;

clustering the non-context tokens into a cluster based upon the frequency vectors according to a lexical correlation among the non-context tokens; and

storing the non-context tokens and a representation of the clusters in a file.

14. The file of claim 13, wherein the clusters may be represented by centroid vectors.

15. The file of claim 13, wherein the predetermined relationship is adjacency.

16. The file of claim 13, wherein the correlation is based on Euclidean distance.

17. The file of claim 13, wherein the correlation is based on Manhattan distance.

18. The file of claim 13, wherein the correlation is based on a maximum distance metric.

19. The file of claim 13, wherein the frequency vectors are normalized based upon the number of occurrences of the non-context token in the corpus.

20. The file of claim 13, wherein the frequency vectors are multi-dimensional vectors, the number of dimensions of which is determined by the number of context tokens and the number of predetermined relationships of non-context tokens to context tokens.